Thermodynamic Prediction of Protein Neutrality

Jesse D. Bloom,[*] [†] Jonathan J. Silberg,[‡] Claus O. Wilke,[†] [§]
D. Allan Drummond,[†] [¶] Christoph Adami,[†] [§] and Frances H. Arnold [*]

[*] Division of Chemistry and Chemical Engineering 210-41,
[†] Digital Life Laboratory 136-93,
[¶] Computation and Neural Systems,
California Institute of Technology,
Pasadena, Californa 91125

[‡] Department of Biochemistry and Cell Biology
Rice University
Houston, TX 77005

[§] Keck Graduate Institute,
535 Watson Drive,
Claremont, California 91711

Corresponding authors:
Jesse Bloom: bloom@caltech.edu
Frances Arnold: frances@cheme.caltech.edu
California Institute of Technology
Mail Code 210-41
Pasadena, CA 91125
Fax: 626-568-8743

In press in *Proc. Natl. Acad. Sci.*
Classification: Biological Sciences, Biochemistry
Total pages including this one: 14
Total number tables: 3
Total number of figures: 4
Total number of Supporting Information tables: 1
Total number of Supporting Information figures: 1
Words in abstract: 147
Characters in paper: 43742

Abstract

We present a simple theory which uses thermodynamic parameters to predict the probability that a protein retains the wildtype structure after one or more random amino acid substitutions. Our theory predicts that for large numbers of substitutions the probability that a protein retains its structure will decline exponentially with the number of substitutions, with the severity of this decline determined by properties of the structure. Our theory also predicts that a protein can gain extra robustness to the first few substitutions by increasing its thermodynamic stability. We validate our theory with simulations on lattice protein models and by showing that it quantitatively predicts previously published experimental measurements on subtilisin and our own measurements on variants of TEM1 β-lactamase. Our work unifies observations about the clustering of functional proteins in sequence space, and provides a basis for interpreting the response of proteins to substitutions in protein engineering applications.

Introduction

The ability to predict a protein's tolerance to amino acid substitutions is of fundamental importance in understanding natural protein evolution, developing protein engineering strategies, and understanding the basis of genetic diseases. Computational and experimental studies have demonstrated that both protein stability and structure affect a protein's tolerance to substitutions. Simulations have shown that more stable proteins have a higher fraction of folded mutants [1, 2, 3, 4] and that some structures are encoded by more sequences than others [5, 6, 7]. Experiments have demonstrated that proteins can be extremely tolerant to single substitutions; for example, 84% of single-residue mutants of T4 lysozyme [8] and 65% of single-residue mutants of *lac* repressor [9] were scored as functional. For multiple substitutions, the fraction of functional proteins decreases roughly exponentially with the number of substitutions, although the severity of this decline varies among proteins [10, 11, 12]. Protein mutagenesis experiments have also underscored the contribution of protein stability to mutational tolerance by finding "global suppressor" substitutions that buffer a protein against otherwise deleterious substitutions by increasing its stability [13, 14].

We unify these diverse experimental and computational results into a simple framework for predicting a protein's tolerance to substitutions. A fundamental measure of this tolerance is the fraction of proteins retaining the wildtype structure after a single random substitution, often called the neutrality [15]. We extend this concept to multiple substitutions by defining the m-neutrality as the fraction of proteins that fold to the wildtype structure among all sequences that differ from the wildtype sequence at m residues. Since mutants that fail to fold also generally fail to function, the m-neutrality provides an upper bound to the fraction of proteins with m substitutions that retain biochemical function. We show that a protein's m-neutrality can be accurately predicted from measurable thermodynamic parameters, and that these predictions capture the contributions of both stability and structure to determining a protein's tolerance to substitutions.

Methods

Lattice Protein Model

We performed simulations with lattice proteins [16] of length $L = 20$ monomers of 20 types corresponding to the natural amino acids. We folded the proteins on a two-dimensional lattice, allowing them to occupy any of the 41,889,578 possible compact or non-compact conformations. The energy of a conformation $\mathcal{C}$ is the sum of the nonbonded nearest-neighbor interactions,

$$E\left(\mathcal{C}\right) = \sum_{i=1}^{L}\sum_{j=1}^{i-2} C_{ij}\left(\mathcal{C}\right) \times \epsilon\left(\mathcal{A}_i, \mathcal{A}_j\right),$$

where $C_{ij}\left(\mathcal{C}\right)$ is one if residues i and j are nearest neighbors in conformation $\mathcal{C}$ and zero otherwise, and $\epsilon\left(\mathcal{A}_i, \mathcal{A}_j\right)$ is the interaction energy between residue types $\mathcal{A}_i$ and $\mathcal{A}_j$, given by Table 5 of [17].

The primary advantage of using lattice proteins is that we can exactly compute the stability of a conformation $\mathcal{C}_t$ as

$$\Delta G_f\left(\mathcal{C}_t\right) = E\left(\mathcal{C}_t\right) + T\ln\left\{Q\left(T\right) - \exp\left[-E\left(\mathcal{C}_t\right)/T\right]\right\},$$

where $Q\left(T\right)$ is the partition sum

$$Q\left(T\right) = \sum_{\{\mathcal{C}_i\}} \exp\left[-E\left(\mathcal{C}_i\right)/T\right]$$

over all conformations, made tractable by noting that there are only 910,972 unique contact sets. All simulations were performed at a reduced temperature of $T = 1.0$

TEM1 β-Lactamase Mutant Libraries

To examine the effects of mutations on the retention of protein function, we constructed mutant libraries of wildtype and the thermostable M182T variant of TEM1 β-lactamase. The 861 bp genes (a kind gift from Brian Shoichet [18]) were subcloned into the pMON:1A2 plasmid [19] with *SacI* and *HindIII* using PCR primers 5'-GCGGCGGAGCTCATGAGTATTCAACATTTCCGTGTCGC-3' and 5'-GCGGCGAAGCTTTTACCAATGCTTAATCAGTGAGGCAC-3' (restriction sites are underlined). We first created a control unmutated library by cutting the gene directly from the plasmid. This unmutated gene was used as the template for a round of error-prone PCR with 100 μl reactions containing 3 ng of template, 0.5 μM of each of the above primers, 7 mM $MgCl_2$, 75 μM $MnCl_2$, 200 μM of dATP and dGTP, 500 μM of dTTP and dCTP, 1X Applied Biosystems PCR buffer without $MgCl_2$, and 5 U of Applied Biosystems *Taq* DNA polymerase. The PCR conditions were 95^oC for 5 minutes, and then 14 cycles of 30 s each at 95^oC, 50^oC, and 72^oC. The product from this PCR was digested with *SacI*/*HindIII* and gel purified, and then used as the template for another identical round of error–prone PCR. This process was repeated to create five libraries with increasing numbers of mutations, which we labeled EP-0 (for the unmutated control) to EP-5 (for the product of the fifth round of error–prone PCR). We quantified the number of doublings for each round by running PCR product versus a known standard on an agarose gel, and found that our protocol consistently yielded ten doublings.

To measure the fraction of genes in the mutant libraries that still encoded functional proteins, we ligated the genes into the pMON:1A2 plasmid with T4 Quick DNA Ligase in 20 μl reactions containing 50 ng each of gene and plasmid, and then transformed 5 μl of the ligation reactions into 50 μl of XL1-Blue Supercompetent cells from Stratagene. The transformed cells were plated on LB-agar plates containing 10 μg/ml of kanamycin (selective only for plasmid) and on LB-agar plates containing 10 μg/ml of kanamycin and 20 μg/ml of ampicillin (selective for both plasmid and active TEM1 gene) at a density that gave 100-300 colonies per unselected plate. The fractions functional were computed as the average of at least five pairs of selected/unselected plates, and are shown in Table 2.

TEM1 mutation frequencies.

Base pairs sequenced	22,800
Total mutations	172
Total AA substitutions	120
Mutation frequency (%)	0.75 ± 0.06
Mutations per gene	6.5 ± 0.5
AA substitutions per gene	4.5 ± 0.4
Mutation types (%)	
A →T, T →A	22
A →C, T →G	9
A →G, T →C	42
G →A, C →T	20
G →C, C →G	1
G →T, C →A	3
frameshift	3

Table 1: Mutation frequencies for TEM1 β-lactamase mutagenesis determined by sequencing 20 unselected clones each from the round five wildtype and M182T error-prone PCR libraries.

To test the ability of our theory to predict the decline in m-neutrality,

The mutation frequency in the round five library was determined by sequencing the first 570 bp of twenty genes each from the unselected wildtype and M182T plates with the sequencing primer 5'-GGTCGATGTTTGATGTTATGGAGC-3'. The wildtype and M182T genes were mutated under identical conditions, and the sequencing found the same nucleotide mutation frequencies for both (0.77 ± 0.08% for wildtype and 0.74 ± 0.08% for M18T2, corresponding to 6.6 ± 0.7 and 6.4 ± 0.7 nucleotide mutations per 861 bp gene). For better statistics, the sequencing results for both libraries were combined to give the data in Table 1. No biases in the locations of the mutations were observed. Eleven mutations occurred twice, which is in good agreement with the expectation of eight duplicate mutations if all possible mutations were equiprobable. The per-round mutation frequency was calculated as 0.15 ± 0.03% (1.3 ± 0.3 nucleotide mutations per gene) by assuming that each round of error-prone PCR introduced the same average number of mutations. To confirm this assumption, we sequenced ten unselected clones each from the wildtype and M182T round one libraries, and found mutation frequencies of 0.16 ± 0.05% for wildtype and 0.19 ± 0.06% for M182T. Standard errors were computed assuming Poisson sampling statistics. More detailed sequencing information is in Table 4 of the Supporting Information.

Results

Thermodynamic Framework for Predicting Neutrality

A protein's native structure is thermodynamically stable [20, 21], with typical free energies of folding (ΔG_f) between -5 and -15 kcal/mol [22]. A mutant sequence folds to the wildtype structure only if the stability of that structure meets some minimal threshold. We call the extra stability of the native structure beyond this minimal threshold $\Delta G_f^{\text{extra}}$ and note that functional proteins always have $\Delta G_f^{\text{extra}} \leq 0$. We define protein's m-neutrality as the fraction of sequences with m substitutions that still meet the stability threshold.

A substitution causes a stability change of

$$\Delta\Delta G = \Delta G_f^{\text{mut}} - \Delta G_f^{\text{wt}}$$

where ΔG_f^{wt} and ΔG_f^{mut} are the wildtype and mutant protein stabilities. Substitutions tend to be destabilizing: although there are no large collections of $\Delta\Delta G$ measurements for truly random substitutions, in a likely-biased collection of more than 2,000 measured $\Delta\Delta G$ values for single-residue substitutions [23], the mean is 0.9 kcal/mol and the values at the 10th and 90th percentiles are -1.0 and 3.2.

The thermodynamic effects of most substitutions are approximately additive [24, 25, 26], meaning that if the stability changes due to two different single substitutions are $\Delta\Delta G^a$ and $\Delta\Delta G^b$, then the stability change due to both substitutions is approximately $\Delta\Delta G^a + \Delta\Delta G^b$. If we know the probability distribution $p_1(\Delta\Delta G)$ that a single random substitution causes a stability change of $\Delta\Delta G$, and if we assume that substitutions are additive, then the net effect $\Delta\Delta G^m$ of m random substitutions is just the sum of m random variables from the probability distribution $p_1(\Delta\Delta G)$. Under this additivity assumption, we can therefore directly calculate the distribution $p_m(\Delta\Delta G^m)$ for $\Delta\Delta G^m$ by performing an m-fold convolution [27] of $p_1(\Delta\Delta G)$.

The m-neutrality $P_f(m)$ is simply the the probability that $\Delta\Delta G^m$ is not more destabilizing than the extra stability $\Delta G_f^{\text{extra}}$ of the wildtype sequence, and can be written as

$$P_f(m) = \int_{-\infty}^{-\Delta G_f^{\text{extra}}} p_m(\Delta\Delta G^m)\,\mathrm{d}(\Delta\Delta G^m). \tag{1}$$

This formula gives a protein's m-neutrality in terms of its extra stability and the distribution of $\Delta\Delta G$ values for all possible single substitutions.

Lattice Proteins Support Predictions

We tested the ability of this simple framework to predict the fraction of lattice proteins that retained the original structure after random amino acid substitutions. Lattice proteins are highly simplified models of proteins that provide a useful tool for studying protein folding [28, 29, 30, 31] and evolution [16, 32] (some example lattice proteins are shown in Figure 1). We can easily measure the m-neutralities of the lattice proteins by making random amino acid substitutions and seeing if the sequences still have $\Delta G_f \leq 0.0$. We can also use Eq. 1 to directly predict the m-neutralities since we can exactly compute ΔG_f and $\Delta\Delta G$ values.

Eq. 1 accurately predicted the m-neutralities of all of the lattice proteins we tested. Lattice proteins with different structures have different m-neutralities, even when they have the same ΔG_f (Figure 1). The 1-neutralities of proteins with different structures and the same ΔG_f look similar, but for larger values of m some proteins clearly show higher m-neutralities than others. For large m, the m-neutralities of all of the proteins converge to a simple exponential of the form

$$P_f(m) \propto \langle \nu_{\text{aa}} \rangle^m$$

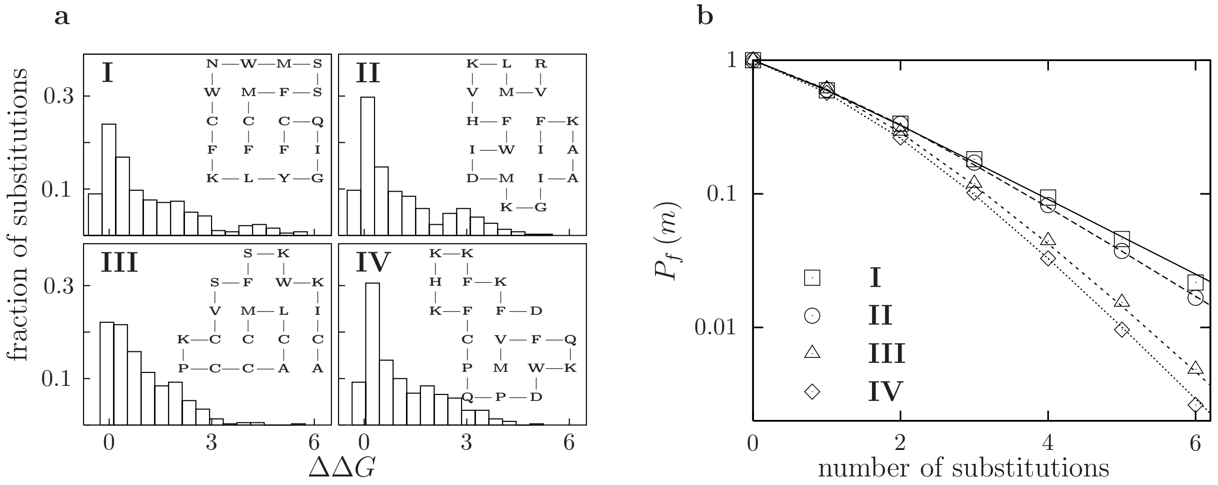

Figure 1: Lattice proteins with different structures but the same stability ($\Delta G_f = -1.0$) converge to different exponential declines in m-neutrality. **(a)** The distributions of $\Delta\Delta G$ for all 380 single amino acid substitutions to the inset lattice proteins. **(b)** The measured (symbols) and predicted (lines) m-neutralities for the four proteins. Proteins are considered folded if $\Delta G_f \leq 0.0$ for the original native structure. The proteins used for the m-neutrality analyses were generated by adaptive walks from random starting sequences, followed by 2.5×10^5 generations of neutral evolution with a population size of 100 and a per generation per residue substitution rate of 5×10^{-5}, selecting for sequences with $\Delta G_f \leq -1.0$ and then taking the first sequence generated with a stability within 0.025 of -1.0. The m-neutralities were computed by sampling all mutants for $m \leq 2$ or 5×10^5 random mutants for $m > 2$. The predicted m-neutralities were computed according to Eq. 1 by numerically convolving the distribution of single–substitution $\Delta\Delta G$ values using generating functions [27] computed with fast–Fourier transforms and a bin size of 0.01.

where $\langle \nu_{\mathrm{aa}} \rangle$ is the average fraction of proteins that are destabilized by a further single random amino acid substitution after several substitutions have already occurred. The underlying reason for the exponential form of this decline is clear: after several substitutions the distribution of ΔG_f among the remaining functional sequences reaches a steady state and each new substitution pushes the same fraction of proteins beyond the stability threshold. The average neutrality $\langle \nu_{\mathrm{aa}} \rangle$ is therefore actually the 1-neutrality averaged over all stable sequences with the wildtype structure. Although $P_f\,(m = 1)$ is similar for all of the protein structures in Figure 1, the factors that give rise to the different values of $\langle \nu_{\mathrm{aa}} \rangle$ for the different structures are present in the distribution of single mutant $\Delta\Delta G$ values, since it is used to predict the m-neutralities for all values of m.

Figure 2 shows the m-neutralities of proteins with the same structure but different stabilities. After several substitutions, all of the proteins converge to the same value of $\langle \nu_{\mathrm{aa}} \rangle$, suggesting that $\langle \nu_{\mathrm{aa}} \rangle$ is a generic property of a protein's structure. On the other hand, the response of a protein to the first few substitutions depends strongly on its stability, with more stable proteins exhibiting higher initial m-neutrality. The high initial m-neutrality of stable proteins is readily rationalized in terms of the thermodynamic model: substitutions tend to disrupt a protein's structure by pushing its stability below the minimal threshold, but proteins with an extra stability cushion are buffered against the first few substitutions [33]. Proteins that sit on the very margin of the minimal stability threshold exhibit lower 1-neutrality than is predicted by an exponential decline because these proteins are less stable than the average folded protein, and so surviving sequences will tend to be more stable than the wildtype sequence and so be more tolerant to the next substitution.

Real Proteins Support Predictions

Our theory makes two main predictions: first, that the decline in m-neutrality is determined by the $\Delta\Delta G$ values for single amino acid substi-

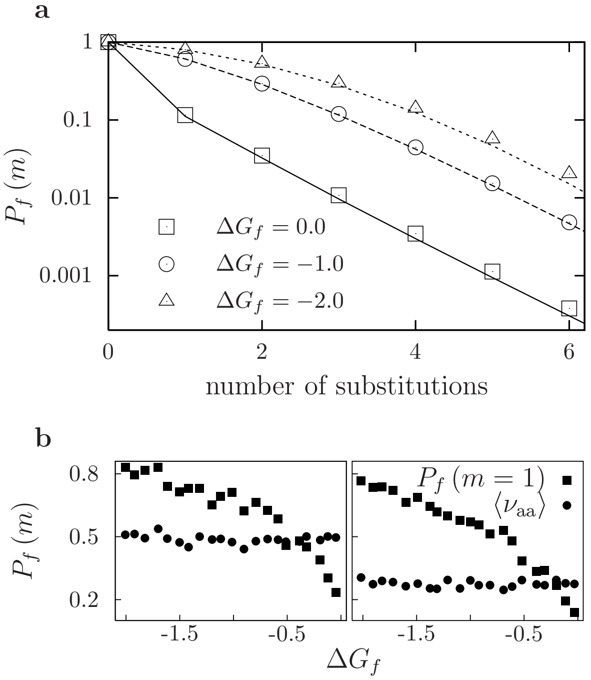

Figure 2: Lattice proteins with the same structure but different stabilities have different 1-neutralities but have the same average neutrality $\langle \nu_{\mathrm{aa}} \rangle$. **(a)** Predicted (lines) and measured (symbols) m-neutralities for proteins with different stabilities and the same structure (**III** in Figure 1). **(b)** Measured values of the 1-neutralities (squares) and average neutralities (circles) for proteins with different stabilities but the same structures (the plots at left and right are for structures **I** and **IV** from Figure 1, respectively). The sequences were generated by finding a sequence with $\Delta G_f = -2.0$ using the procedure described in Figure 1, and then using this sequence as a starting point for neutral evolution selecting for the indicated target stabilities. The proteins with different stabilities are highly diverged, with average pairwise sequence identities of 15% and 41% for the structures at left and right, respectively. The m-neutralities were computed as in Figure 1, and $\langle \nu_{\mathrm{aa}} \rangle$ was computed as the square root of the 6-neutrality divided by the 4-neutrality.

TEM1 mutant library measurements.

Round	$\langle m_{nt} \rangle$	$\langle m_{aa} \rangle$	WT	M182T
0	0.0 ± 0.0	0.0 ± 0.0	0.76 ± 0.03	0.74 ± 0.04
1	1.3 ± 0.2	0.9 ± 0.1	0.59 ± 0.03	0.68 ± 0.03
2	2.6 ± 0.3	1.8 ± 0.2	0.47 ± 0.03	0.54 ± 0.02
3	3.9 ± 0.4	2.7 ± 0.2	0.28 ± 0.02	0.45 ± 0.04
4	5.2 ± 0.4	3.6 ± 0.3	0.18 ± 0.01	0.28 ± 0.01
5	6.5 ± 0.5	4.5 ± 0.4	0.13 ± 0.01	0.20 ± 0.02

Table 2: Measured fractions of functional proteins in mutant libraries of wildtype and the thermostable M182T variant of TEM1 β-lactamase. The table shows the number of rounds of error-prone PCR, the average number of nucleotide mutations per gene, and the fractions of mutated genes that confer ampicillin resistance in *E. coli*. Values are shown ± their standard errors.

tutions, and second, that among proteins with the same structure, more stable variants will have higher m-neutralities. We tested these predictions against measurements of the fractions of functional proteins in mutant libraries of subtilisin and variants of TEM1 β-lactamase. Our theory is designed to predict the fraction of proteins that retain the wildtype structure, but the experiments measure the fraction of proteins that retain function. However, since proteins that fail to fold also generally fail to function, our theory provides an upper bound on the fraction of functional proteins. We expect that for many proteins this upper bound will closely approximate the actual fraction functional since mutagenesis studies suggest that most functionally disruptive random substitutions disrupt the structure rather than specifically affect functional residues [13, 34, 35].

To test the ability of our theory to predict the decline in m-neutrality, we used data on the fractions of functional proteins in subtilisin mutant libraries created by Shafikhani and coworkers [10] (population 6B of Table 2 of [10], normalized by the fraction of functional clones in the control libraries) and our own mutant libraries of TEM1 (Table 2). Each mutant library contains a distribution of sequences with different numbers of amino acid mutations. The form of this distribution is known: the probability that a sequence in a library with an average of $\langle m_{nt} \rangle$ nucleotide mutations created by N cycles of PCR with a PCR efficiency of λ will have m_{nt} mutations is

$$f\left(m_{nt}\right) = \left(1+\lambda\right)^{-N} \sum_{k=0}^{N} \binom{N}{k} \lambda^{k} \frac{(kx)^{m_{nt}} e^{-kx}}{m_{nt}!}$$

where $x = \langle m_{nt} \rangle \left(1+\lambda\right) / \left(N\lambda\right)$ [36, 37]. Subtilisin was mutagenized using 13 PCR cycles with 10 effective doublings [10], so N is 13 times the number of rounds of error–prone PCR and $\lambda = 0.77$. TEM1 was mutagenized using 14 PCR cycles with 10 effective doublings, so N is 14 times the number of rounds and $\lambda = 0.71$. We confirmed that $f\left(m_{nt}\right)$ accurately describes the distribution of mutations in our libraries (Figure 5 of the Supporting Information).

The expected fraction of folded sequences in a mutant library is easily calculated from $f\left(m_{nt}\right)$ and the probability $P_f\left(m_{nt}\right)$ that a sequence is still functional after m_{nt} nucleotide mutations as

$$\mathcal{F} = \sum_{m_{nt}=0}^{\infty} f\left(m_{nt}\right) \times P_f\left(m_{nt}\right).$$

We calculated the probability $P_f\left(m_{nt}\right)$ that a sequence was still folded after m_{nt} nucleotide mutations by using two existing computer programs for estimating the $\Delta\Delta G$ values for single substitutions to proteins with known structures (PDB structure 1IAV for subtilisin and 1BTL for TEM1): Gilis and Rooman's PoP-MuSiC potential [38] and Serrano and coworkers' FOLDEF potential [39] with van der Waals clash energies. Since the genetic code makes nucleotide mutations more likely to induce some amino acid substitutions than others, and since error–prone PCR introduces a non-random distribution of nucleotide mutations, we weighted each $\Delta\Delta G$ value by the probability that it would be induced by a single nucleotide mutation made according to the observed error–prone PCR nucleotide mutation frequencies (given in Table 1 of [10] for subtilisin and Table 1 of the current work for TEM1). We assigned a $\Delta\Delta G$ of zero to synonymous nucleotide mutations since they do not cause an amino acid substitution, and we assigned a $\Delta\Delta G$ of 25 kcal/mol to frameshift and nonsense mutations since premature truncation is expected to inactivate the protein. We ignored the small number of substitutions for which PoPMuSiC failed to calculated a $\Delta\Delta G$. With this weighted $\Delta\Delta G$ distribution for nucleotide mutations, all we needed to construct $P_f(m_{nt})$ according to Eq. 1 was the value of $\Delta G_f^{\text{extra}}$. This cannot be measured directly since we do not know the minimal stability threshold. However, since $\Delta G_f^{\text{extra}}$ only influences the initial behavior of the m-neutrality and does not affect the limiting decline (Figure 2), and since we have six data points for each protein, we could do a least-squares fit of $\Delta G_f^{\text{extra}}$ to the data and still test the ability of the theory to predict the decline in the fraction of functional proteins.

Figure 3 shows the measured fractions of functional proteins for subtilisin and wildtype TEM1 versus the theoretical predictions made with PoPMuSiC and FOLDEF. The theoretical predictions closely match the measured fractions of functional proteins in all cases, with subtilisin exhibiting slightly higher m-neutralities than TEM1.

The second major prediction of our theory is that among proteins with the same structure, more stable variants will exhibit higher initial m-neutralities, but converge to same average neutrality. To test this prediction, we compared the fractions of functional proteins in mutant libraries of wildtype and the M182T variant of TEM1. The M182T variant differs from wildtype by only a single substitution yet is 2.7 kcal/mol more stable [18], so we predict that it should exhibit a higher fraction of functional proteins at the same level of mutation. Figure 4 shows the measured fractions functional for wildtype and the M182T variant, as well as the

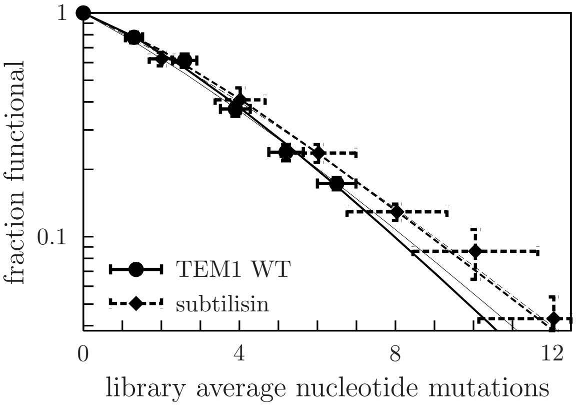

Figure 3: Theoretical predictions and fractions of functional proteins in mutant libraries of subtilisin (dashed lines) and TEM1 β-lactamase (solid lines) genes. Thick lines show predictions made using PoPMuSiC [38] and thin lines show predictions made using FOLDEF [39]. The TEM1 measurements are from Table 2, normalized by the values from the control unmutated library.

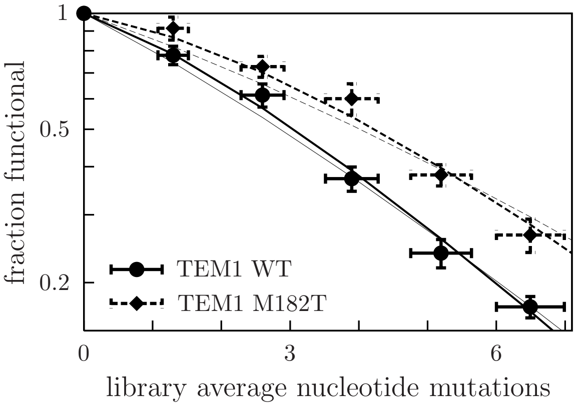

Figure 4: The more stable M182T variant of TEM1 β-lactamase (dashed lines) exhibits a higher fraction of functional mutants relative to wildtype (solid lines), as predicted. Thick lines show predictions made using PoPMuSiC [38] and thin lines show predictions made using FOLDEF [39]. The measurements are from Table 2, normalized by the values from the control unmutated library.

theoretical predictions made using both PoPMuSiC and FOLDEF. As predicted, the M182T variant exhibits a higher fraction of functional proteins, and once again the predictions made with both potentials are in good agreement with the experimental measurements.

To further explore the range of possible neutralities for different proteins, we used $\Delta\Delta G$ values from PoPMuSiC to predict the expected average neutralities to both amino acid substitutions ($\langle \nu_{\rm aa} \rangle$) and nucleotide mutations ($\langle \nu_{\rm nt} \rangle$) for proteins chosen from several different CATH [40] protein structure classifications. Since we do not know $\Delta G_f^{\rm extra}$ for these proteins, we computed the fraction of proteins expected to be inactivated by the 10th mutation since after this many mutations effects due to the initial protein stability should be small. Table 3 shows the predicted average neutralities to both random amino acid substitutions and nucleotide mutations made according to the mutation probabilities of our TEM1 mutagenesis. The predicted average neutralities differ considerably, showing that our theory predicts that different proteins can have substantially different neutralities.

Discussion

We have presented a theory for calculating the probability that a protein will retain its structure after random amino acid substitutions, and have confirmed the main theoretical predictions with simulations and experiments. Our theory naturally separates a protein's m-neutrality into components due to structure and stability. The eventual severity of the exponential decline in m-neutrality with the number of substitutions is a property of a protein's structure. On the other hand, increased stability confers greater tolerance to the first few substitutions, in effect allowing a protein to "take a few hits" before it is pushed into the inevitable structurally determined exponential decline in m-neutrality. This increased tolerance to mutations due to extra stability is probably also the underlying reason for the existence of global suppressor mutations [13, 14] that buffer proteins against otherwise deleterious mutations.

The major assumption underlying our theory is that the thermodynamic effects of substitutions are additive. This assumption is clearly not strictly true since protein residues do interact. Substitutions are most likely to be non-additive if the mutated residues are in close contact in a protein's structure [24, 25]. Since proteins are large, two randomly chosen residues will rarely contact each other, and so although the additivity assumption is certainly violated for some specific combinations of substitutions, it is accurate when averaged over all possible substitutions. When we apply our theory to measurements of the fraction of mutant proteins that retain function we are making a second assumption by ignoring the possibility that some substitutions may disrupt a protein's function in ways other than affecting its stability. Therefore, for proteins with a high fraction of functional residues, our theory provides only an upper bound on the fraction of functional proteins. However, our theory's remarkable success for both the subtilisin and the TEM1 mutant libraries suggests that this assumption is also valid.

Our theory provides a quantitative rationale for earlier work with lattice proteins on the organization of functional proteins in sequence space. Bornberg-Bauer and Chan [2] proposed that proteins are located in superfunnels in sequence space with the most stable sequence having the most neutral neighbors; others have reported that folded proteins surround highly stable prototype sequences in sequence space [41, 3, 4], and Shakhnovich and coworkers [1] showed that proteins with a large energy gap between the lowest and second lowest energy conformations are stabilized against mutations. We provide a clear explanation: more stable proteins are able to tolerate more of the possible mutations before unfolding, and so a higher fraction of their neighboring sequences fold.

In addition to these stability-based effects, different protein structures have different inherent designabilities, with more sequences folding into some structures than others [5, 42, 43].

Predicted Average Neutralities

PDB	Protein	CATH architecture	Length	$\langle \nu_{\rm nt} \rangle$	$\langle \nu_{\rm aa} \rangle$
1IAV	subtilisin	$\alpha\beta$ 3-layer sandwich	269	0.65	0.55
1B9C	GFP	β barrel	236	0.62	0.56
1BTL	TEM1 β-lactamase	$\alpha\beta$ 3-layer sandwich	263	0.58	0.46
1RLV	tRNA endonuclease	not classified	305	0.55	0.44
1HZW	thymidylate synthase	$\alpha\beta$ 2-layer sandwich	290	0.50	0.41
2BNH	ribonuclease inhibitor	$\alpha\beta$ horseshoe	457	0.45	0.35
1HEL	hen lysozyme	α orthogonal bundle	129	0.43	0.38

Table 3: Predictions of the average neutralities of various proteins to both nucleotide mutations ($\langle \nu_{\rm nt} \rangle$) and amino acid substitutions ($\langle \nu_{\rm aa} \rangle$). The codes for the PDB structures and the CATH [40] architectures are shown along with the lengths of the protein chains in the PDB structures (in all cases we consider chain A). The average neutralities are computed by calculating the fraction of sequences predicted be inactivated by the 10th mutation or substitution, using $\Delta\Delta G$ values from PoPMuSiC [38] and assuming that the proteins all have the same value of $\Delta G_f^{\rm extra}$ as wildtype TEM1 β-lactamase. The values of $\langle \nu_{\rm nt} \rangle$ are computed assuming that nucleotide mutations are made according to the error–prone PCR mutation frequencies of Table 1.

Proteins with more designable structures might be expected to show a higher average neutrality since their structures occupy a larger fraction of sequence space. The average neutrality $\langle \nu_{\rm aa} \rangle$ therefore provides a quantitative measure of designability that can be estimated with current computational techniques.

Our work suggests a more nuanced approach to experimentally analyzing protein neutralities than has been applied in the past. Loeb and coworkers [11] have performed a careful analysis of the neutralities of several proteins or regions of proteins under the assumption of a strict exponential decline in m-neutrality. However, our work suggests that a protein's m-neutrality can deviate from a strict exponential for the first few substitutions if the protein has a large amount of extra stability, as we show for the M182T variant of TEM1. Experimental mutagenesis studies suggest that during natural evolution, proteins accumulate mildly destabilizing mutations that are counterbalanced by stabilizing mutations [26]. We suggest that it is also important to examine whether some natural proteins have systematically accumulated stabilizing mutations in order to provide them with additional robustness [15] to amino acid substitutions.

Our work also has applications in protein engineering. Directed evolution involves screening libraries of mutant proteins for new or improved functions [44]. Each round of directed evolution typically introduces only one or two amino acid substitutions because the rapid decline in m-neutrality means that higher mutation rates will yield libraries of mostly unfolded proteins. Our work suggests that using highly stable parents for directed evolution should increase the fraction of folded mutants at a given level of substitutions. It also provides a method for predicting which structures will better tolerate large numbers of substitutions.

Acknowledgements

We thank Brian Shoichet for providing us with genes for the TEM1 β-lactamase variants, and Titus Brown for programming assistance. We thank Michelle Meyer and Eric Zollars for helpful advice and discussions, and two anonymous reviewers for their helpful comments. J.D.B. is supported by a Howard Hughes Medical Institute predoctoral fellowship. D.A.D. is supported by the National Institutes of Health, National Research Service Award 5 T32 MH19138 from the National Institute of Mental Health. C.A. and C.O.W. were supported in part by the National Science Foundation under grant DEB-9981387.

References

[1] Broglia, R. A, Tiana, G, Roman, H. E, Vigezzi, E, & Shakhnovich, E. (1999) *Phys. Rev. Lett.* **82**, 4727–4730.

[2] Bornberg-Bauer, E & Chan, H. S. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 10689–10694.

[3] Wingreen, N. S, Li, H, & Tang, C. (2004) *Polymer* **45**, 699–705.

[4] Xia, Y & Levitt, M. (2004) *Proteins* **55**, 107–114.

[5] Li, H, Helling, R, Tang, C, & Wingreen, N. (1996) *Science* **273**, 666–669.

[6] Shakhnovich, E. I & Gutin, A. M. (1991) *J. Theor. Biol.* **149**, 537–546.

[7] Govindarajan, S & Goldstein, R. A. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 3341–3345.

[8] Rennell, D, Bouvier, S. E, Hardy, L. W, & Poteete, A. R. (1991) *J. Mol. Biol.* **222**, 67–87.

[9] Markiewicz, P, Kleina, L. G, Cruz, C, Ehret, S, & Miller, J. H. (1994) *J. Mol. Biol.* **240**, 421–433.

[10] Shafikhani, S, Siegel, R. A, Ferrari, E, & Schellenberger, V. (1997) *BioTechniques* **23**, 304–310.

[11] Guo, H. H, Choe, J, & Loeb, L. A. (2004) *Proc. Natl. Acad. Sci. USA* **101**, 9205–9210.

[12] Daugherty, P. S, Chen, G, Iverson, B. L, & Georgiou, G. (1999) *Proc. Natl. Acad. Sci. USA* **97**, 2029–2034.

[13] Shortle, D & Lin, B. (1985) *Genetics* **110**, 539–555.

[14] Poteete, A. R, Rennell, D, Bouvier, S. E, & Hardy, L. W. (1997) *Protein Sci.* **6**, 2418–2425.

[15] Wilke, C. O & Adami, C. (2003) *Mutat. Res.* **522**, 3–11.

[16] Chan, H. S & Bornberg-Bauer, E. (2002) *Applied Bioinformatics* **1**, 121–144.

[17] Miyazawa, S & Jernigan, R. L. (1985) *Macromolecules* **18**, 534–552.

[18] Wang, X, Minasov, G, & Shoichet, B. K. (2002) *J. Mol. Biol.* **320**, 85–95.

[19] Meyer, M. M, Silberg, J. J, Voigt, C. A, Endelman, J. B, Mayo, S. L, Wang, Z. G, & Arnold, F. H. (2003) *Protein Sci.* **12**, 1686–1693.

[20] Anfinsen, C. B. (1973) *Science* **181**, 223–230.

[21] Ellis, R. J, Dobson, C. M, & Hartl, U. (1998) *Trends Biochem. Sci.* **23**, 468–468.

[22] Fersht, A. R. (1999) *Structure and Mechanism in Protein Science.* (W. H. Freeman and Company, New York).

[23] Bava, K. A, Gromiha, M. M, Uedaira, H, Kitajimi, K, & Sarai, A. (2004) *Nucleic Acids Res.* **32**, D120–D121.

[24] Wells, J. A. (1990) *Biochemistry* **29**, 8509–8517.

[25] Zhang, X. J, Baase, W. A, Shoichet, B. K, Wilson, K. P, & Matthews, B. W. (1995) *Protein Eng.* **8**, 1017–1022.

[26] Serrano, L, Day, A. G, & Fersht, A. R. (1993) *J. Mol. Biol.* **233**, 305–312.

[27] van Kampan, N. G. (1992) *Stochastic Processes in Physics and Chemistry.* (Elsevier).

[28] Dill, K. A, Bromberg, S, Yue, K, Fiebig, K. M, Yee, D. P, Thomas, P. D, & Chan, H. S. (1995) *Protein Sci.* **4**, 561–602.

[29] Hinds, D. A & Levitt, M. (1994) *J. Mol. Biol.* **243**, 668–682.

[30] Shakhnovich, E. I & Gutin, A. M. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 7195–7199.

[31] Socci, N. D, Onuchic, J. N, & Wolynes, P. G. (1998) *Proteins* **32**, 136–158.

[32] Xia, Y & Levitt, M. (2004) *Curr. Opin. Struct. Biol.* **14**, 202–207.

[33] Axe, D. D. (2004) *J. Mol. Biol.* **341**, 1295–1315.

[34] Loeb, D. D, Swanstrom, R, Everitt, L, Manchester, M, Stamper, S. E, & III, C. A. H. (1989) *Nature* **340**, 397–400.

[35] Pakula, A. A, Young, V. B, & Sauer, R. T. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 8829–8833.

[36] Sun, F. (1995) *J. Comput. Biol.* **2**, 63–86.

[37] Drummond, D. A & Arnold, F. H. (2004) *ArXiv:q-bio.QM/0411041.*

[38] Gilis, D & Rooman, M. (2000) *Protein Eng.* **13**, 849–856.

[39] Guerois, R, Nielsen, J. E, & Serrano, L. (2002) *J. Mol. Biol.* **320**, 369–387.

[40] Orengo, C. A, Michie, A. D, Jones, S, Jones, D. T, Swindells, M. B, & Thornton, J. M. (1997) *Structure* **5**, 1093–1108.

[41] Taverna, D. M & Goldstein, R. A. (2002) *J. Mol. Biol.* **315**, 479–484.

[42] Wolynes, P. G. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 14249–14255.

[43] England, J. L & Shakhnovich, E. I. (2003) *Phys. Rev. Lett.* **90**, 218101.

[44] Arnold, F. H. (1998) *Acc. Chem. Res.* **31**, 125–131.

Mutations in the round five TEM1 libraries

Clone	TEM1 WT	TEM1 M182T
1	G43A (A15T), T64C (F22L), C68A (A23D), A90G, G154A (G52S), A375C, T407C (L136P), A426G, T479G (L160R)	G122T (R41L), A158T (K53M), A185C (E62A), G198A (M66I), C204T, T246C, T408C
2	T52C (C18R), A71C (H24P), A149G (N50S), T200A (M67K), A319C (T107P), T332A (L111H), A388T (N130Y), T393C	A94G (K32E), T161A (I54N), A195T, G385A (D129N), A430T (STOP), A467G (H156R), T470A (V157E)
3	C28A (L10I), delT30, T38C (F13S), T50C (F17S), A123T, A151G (S51G), T190C (F64L), G327C (K109N)	T49C (F17L), T50C (F17S), A120G, delG128, C249T, T278C (I93T)
4	T42C, T200G (M67R), T384A (S128R), A423G, G459T (M153I), A503C (N168T), delT504	T33A, T130C (Y44H), A264G, G268A (G90S), C466T (H156Y), C545T (A182V)
5	A77G (E26G), G100A (A34T), A149G (N50S), C177T, T236A (STOP), A418T (I140F)	A90G, T223A (C75S), C241T (R81C), A263G (Q88R), T339A (D113E), T410A (L137Q)
6	T171G (S57R), A186G, T317C (V106A), T465C, A518G (N173S), A549G	T161C (I54T), T243C, A364G (S122G), C427T (P143S), A550G (M184V)
7	C27T, T52A (C18S), G67A (A23T), T216C, T374C (I125T), A491G (E164G)	T69A, T174C, A195G, G488A (STOP), A559G (T187A)
8	T39C, T40C (F14L), T51A (F17L), T361A (C121S), A472G (T158A), G476A (R159H)	T41C (F14S), T302A (V101D), A467G (H156R), G476A (R159H), A482G (D161G)
9	A158T (K53M), A186G, G306A, T384G (S128R), T539G (M180R)	T144C, C276T, A281G (H94R), A513G (I171M)
10	T41A (F14Y), T86A (V29E), A464G (D155G), A570G	A148T (N50Y), C153T, A199G (M67V), C378T
11	delA7, T245A (V82D), T332C (L111P)	C70T (H24Y), T245G (V82G), T286A (S96T), A357G
12	A95G (K32R), A308T (Y103F), G525A	T33C, C147T, A348G, T404C (L135S)
13	T447G (F149L), A452G (H151R), A513G (I171M)	A95G (K32R), T225A (STOP), G383A (S128N), C416T (T139M)
14	A321G, A524G (E175G), A538T (M180L)	A258C (Q86H), C267T, A315T
15	T112C, A167G (E56G), G566A (R189H)	T173A (F58Y), T393C, C539T (T180M)
16	T112C, A158G (K53R)	G340A (G114S), C436T, A454G (N152D)
17	C101G (A34G), T384A (S128R)	A157T (STOP), T403G (L135V), C440T (T147I)
18	A158G (K53R), T233A (V78E)	delT42, T380A (M127K)
19	C28T (L10F), A516C	G168T (E56D), C187T (R63C)
20		A352T (STOP)

Table 4: The mutations found in the 20 unselected clones from the round five wildtype and M182T TEM1 β-lactamase libraries. For nonsynonymous mutations, the amino acid change is shown in parentheses.

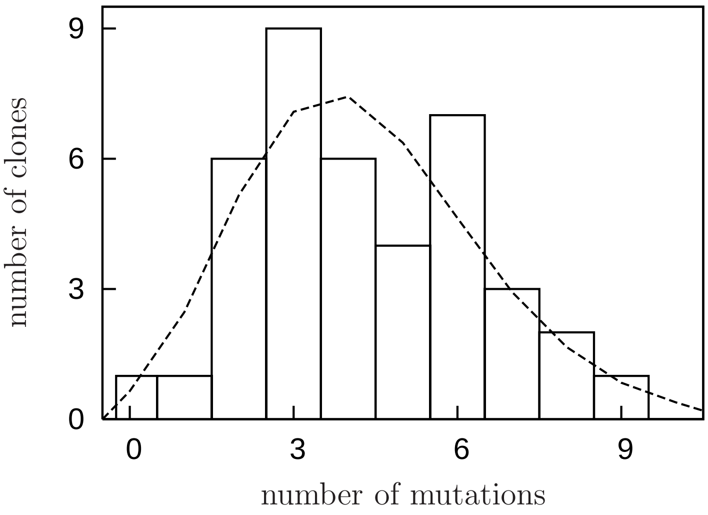

Figure 5: The observed distribution of mutations among the 40 sequenced unselected TEM1 β-lactamase clones from the round-five mutant libraries (bars) agrees with the theoretical predictions (lines) for the distribution of mutations in an error–prone PCR library. The theoretical predictions are made with the equation $f\left(m_{nt}\right)$ described in the text, with an average number of nucleotide mutations of $\langle m_{nt} \rangle = 4.3$ for the 570 bp sequenced region of the 861 bp gene. A chi-square test demonstrated that the observed distribution is consistent with the theoretical predictions, with a P-value > 0.9 that a difference at least this large between the observed and predicted values would occur by chance.